Text Book on

Fundamentals of Multimedia Computing

Authors:
Gerald Friedland
and
Ramesh Jain

Draft for Comments

Chapter X: Multimedia Information Retrieval

INTRODUCTION

In a very short time period, Multimedia Information Retrieval (MIR) has rapidly emerged as a technology that is needed to solve many useful applications faced by people in different aspects of their regular activities. Serious efforts in this area could be traced to early 1990s. After a slow start, this area became very active and attracted researchers from several areas of computer science, including computer vision, databases, and information retrieval. Devices like digital cameras and phone cameras combined with progress in compression and availability of bandwidth brought a major change in the lifestyles of people first in advanced countries and then everywhere in the world. In fact, the rapid progress in technology created strong demand for organization and access to multimedia data, but the techniques for multimedia information retrieval have been slower to develop than the volume of data. The basic problem in MIR system is connecting different types of data sources to users with diverse background and different needs, as shown in Figure 1.

Multimedia Information Retrieval (MIR) contains three important components: Multimedia, Information, and Retrieval. Each of these terms is important and should be clearly understood to understand MIR.

While thinking of the MIR problem, it is natural to think that increase in number of types of data such as images, text, and audio will result in increased complexity of organization, indexing, and retrieval. Contrary to this obvious thought, however, by adopting a right perspective and using opportunistic information from disparate sources, the availability of correlated and complementary multimedia data simplifies the problems significantly.

It is important to emphasize that most, if not all, techniques in MIR are related to information that will be derived by and for humans, but the processing is done mostly by computers. This fact is important because MIR is naturally influenced by IR (Information Retrieval) which deals with text. In text most successes are because the information retrieved is directly available in data, while the information of interest in sensory data requires processing to extract this information. Whenever IR systems try to retrieve information not directly available in data, they also face challenging problems. A popular problem often discussed in MIR community, the semantic gap, is directly related to this issue.

Retrieval is usually interpreted as the operation of accessing information from human or computer memory. In many cases retrieval, query, and search are used to mean similar things. Clearly there are subtle differences and that is the reason that database

community uses query; text search popularized information retrieval; and the Web community likes to search. All of these terms are related to finding appropriate information in some application context from a large volume of data in memory. Depending on the sources of data and the application context, sometimes one is interested in precise answers that are directly available in the data; sometimes in information that is derived from the data; and in other cases just finding related sources that may contain information. These cases have different scope and require different techniques. Another important fact is that in most applications, retrieval is one step in the overall solution and the application context influences the requirements from this step significantly.

Most of the knowledge in the world is captured and stays in the form of experiences in the data of different sensing modalities. Our current technology tamed knowledge in text because in text the audio data is converted to symbols by humans. Text is a visual representation of a subset of sounds created using the human vocal chord. Since this was the only technology to make human experience widely sharable and usable, text has been popular for more than 5000 years. Text became particularly dominant after Gutenberg invented moveable print. Today, even the WWW is dominated by text. Photos and videos are increasing on the Web, but are organized and usually accessed on WWW using tags and keywords.

Humans are extremely adept in dealing with sensory and symbolic information by effortlessly converting sensory information to symbolic form and processing this hybrid form of information effortlessly. Converting sensory data to symbols in computer systems has attracted significant research, but so far, has proven exceedingly difficult. A primary difficulty in developing computational techniques for automated sensory interpretation lies in our inability to formally represent and effectively model the appropriate context within which the sensory information should be interpreted. A specific impediment in using contextual information has been in our inability to constrain the scope of the potentially infinite number of uncertain and imprecise context-defining variables. By organizing all multimedia experiences and making them as searchable using experiences as modern search engines have done using keywords, tremendous progress will be made in inductive reasoning to generalize and create new knowledge, and abductive reasoning to verify facts from data.

In this chapter, we will present basic concepts and techniques related to accessing multimedia data. We will start with the structured data in databases, discuss information retrieval to deal with accessing information in text, and then present techniques developed and being explored in MIR.

Multimedia computing addresses a problem that many other fields like computer vision, databases, and information retrieval face: connecting data and users. As shown in the figure below, data exists in many forms, ranging from bits to alphanumeric documents to photos and video. On the other hand users of the data in a modern computing environment may come from many different education backgrounds, of different culture, and of different socio-economic status. The challenge is how to connect a user with a data source so the user can use the data he needs to solve his application. A key point to

remember is that a user is never interested in what and where the data is; she is only interested in solving the problem at hand. The major hurdle in connecting users to the data is often referred to as the semantic gap. This is explained in more details in Chapter on Context.

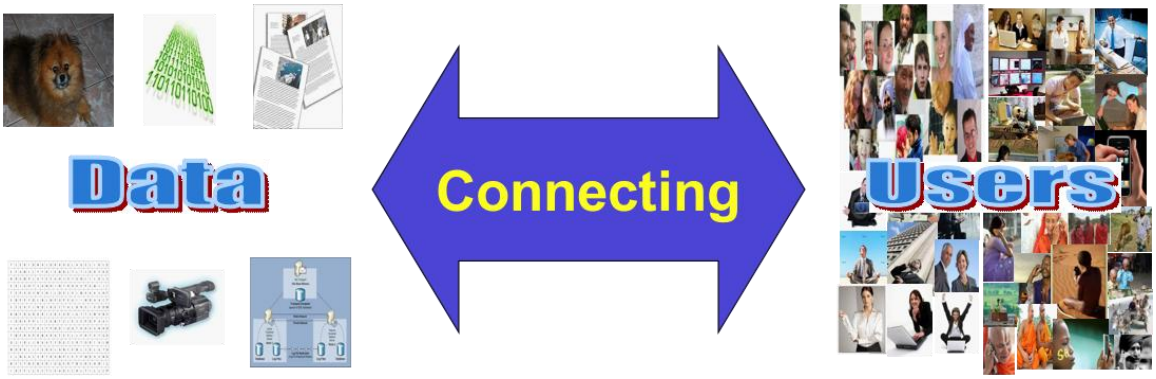

Figure 1: Multimedia data could be in any form ranging from text to signals like video. People from very diverse background may come to a repository in a MIR system to access the data to satisfy their needs.

Structured and Un-structured data

Nature of data plays a very important role in how it could be organized and retrieved. Some data could be in very structured format such that the sequence of data items is well defined and is known to the system. In such a case, system knows how to interpret data and can store it using indexing techniques for efficient retrieval of the data. At the other extreme is the data where all data appears in a form that is not known to the system and may contain data items or elements in random types and order. In such a case the system cannot use any indexing schemes and usually stores the data based on the name of the file and some other meta data such as the date of creation of the file containing the data.

Clearly, when you store something—data or physical things—usually the goal is to be able to access them later. All storage techniques follow organization principles to let us retrieve information rapidly and efficiently. Computer scientists were the first to develop powerful data models to organize and store data so that we could retrieve it quickly at the right time with a minimal amount of computing effort. When data started to grow, we developed database techniques to deal with structured data. Relational databases became popular because they let us map data on storage devices and articulate queries in a unified manner. As a result, developers created applications that took advantage of relational databases. Although object-oriented databases offered a more flexible data model, they weren't successful because of their poor efficiency and efficacy compared to the relational model. To work around this problem, people learned to represent object-oriented models on relational databases. The Internet changed things by allowing millions of people to create Web sites. These Web sites mostly contained text documents. When the Internet started to grow rapidly, portals appeared on the scene with their arsenal of

keyword-based searches and taxonomies inspired by library science approaches. This worked initially, but we realized that keywords couldn't capture the meaning of documents. So, techniques based on artificial intelligence approaches to natural-language understanding or on neural networks approaches to classify documents started cropping up. It became clear that all these approaches would have only limited success and a limited lifespan.

Since there was a clear need to access increasing volume of text data, commonly considered semi-structured data, people developed techniques that could help in providing some structure to this new common type of data and introduced a new approach structure using Extensible Markup Language (XML). . XML uses a different approach to structuring data by asking document creators to introduce enough clues, or structure, in the document so that an automatic process can read what the document or a section of it is about. This metadata approach enables advanced systems to know more about the document than today's automatic techniques can. It also has the ability to work gracefully with more automation.

XML introduces structure in otherwise unstructured documents. That is, it structuralizes text. Multimedia data, like other data, must be stored using organization principles that will help enable management and retrieval. Moreover, multimedia data should be organized more carefully because of its time–serial nature and its enormous size. Another difficulty is that current metadata for audio, video, images, and other similar sources is more about the data than about its semantic content. The tags in XML introduce semantic partitioning of text. Techniques for introducing the semantic partitioning of video, audio, and images are needed. Multimedia researchers have spent considerable effort on developing automatic techniques for video and audio segmentation and for indexing images based on some basic characteristics such as color and texture. These techniques are very useful and will revolutionize how we'll organize multimedia data someday in the future. However, we need to organize multimedia data today. The current automatic techniques for semantic partitioning are even more infantile than those for text. The only solution may be to develop powerful approaches for structuralizing multimedia data, which could prove to be as revolutionary as the introduction of XML.

One can not provide organization and access to large volume of data without using some structure in the data. In some cases, such as relational databases, the data is created and made available to the system in a well-structured format. In other cases when the data is not directly entered in the system, some techniques must be used to identify the structure that will help in providing the functionality for organizing and accessing the data. Multimedia information retrieval techniques, discussed in this chapter, are all about defining structure that must be used to organize and access the data and applying techniques for data analysis that will help in extracting this information from the data and storing it.

Databases

Searching for data is an old problem. When computers started becoming popular, and started finding applications in businesses as well as in our applications involving large volumes of data, it became important to develop systems that will help in organization,

storage, management, and retrieval of data (OSMR). After early navigational and network databases, the relational model introduced by Edgar Codd became popular and became the foundation for most of the commercial database systems. The uniform set-theoretic representation of entities and their attributes allowed efficient OSMR operations on large volumes of data. The basic data structure behind relational data model is the concept of a Table. All information in relational databases is stored in tables in which each row represents an entity and columns represent their attributes.

A very important concept commonly used in databases is that there are three distinct levels, shown in Figure 2, at which data must be viewed and managed:

1. **Physical:** At this level the system takes care of storing all data and takes care of accessing, adding, deleing, and updating any particular records that are rows in a table, without other levels having to worry where it is really stored in a storage device.
2. **Logical:** This level represents how the data is modeled by application designers. Thus at this level the system knows what are the entities and their attributes that will be used by the database. Using techniques like Entity-Relationship models, the logical model is designed and then translated to tabular representations in the database. At this level the system is only aware of the models of objects (entities) and their characteristics (attributes) that are used in the system. This level is designed based on the desired functionality that a particular application must have.
3. **View:** A database is used by many different types of users, each of which must have different rights and privileges to access different type of data. Thus, in a university database, an instructor may have access to look at the grades of a every student in his class and modify those, but a student can only see his/her grade and will not be able to modify those. Not all functionality of a database is exposed and made available to every user. Different users see only a subset of data and have different rights with respect to addition, deletion, and updating of the data. Each user thus has just *a view* of the database.

An operation to access information from a database is commonly called a query. A query in a relational database is articulated using a sentence in SQL (Structured Query Language).

Information Retrieval and Search

Information retrieval is a field that addresses techniques for searching for information in documents. In early stages it was predominantly concerned with issues to find information in books. As the technology evolved and books started becoming electronic, the field evolved to consider searching for information in a collection of books. Information retrieval as a field was limited in applications until the emergence of the World Wide Web. With the arrival of the Web, each node on this Web was a document and all these documents were connected on the Web. In a sense one could consider that the Web was a giant document of documents. To search for information on the Web, it was important to have technology that will allow efficient organization and retrieval of

information. This need resulted in significant development of techniques in information retrieval. However, the nature of the Web is significantly different in many dimensions from traditional documents and techniques for searching information on Web must be designed considering specific needs of the Web. Search techniques on the Web utilize concepts from information retrieval but have been emerging to address several other aspects. In this section, we will discuss basic concepts from information retrieval and how some of those are being applied in search approaches.

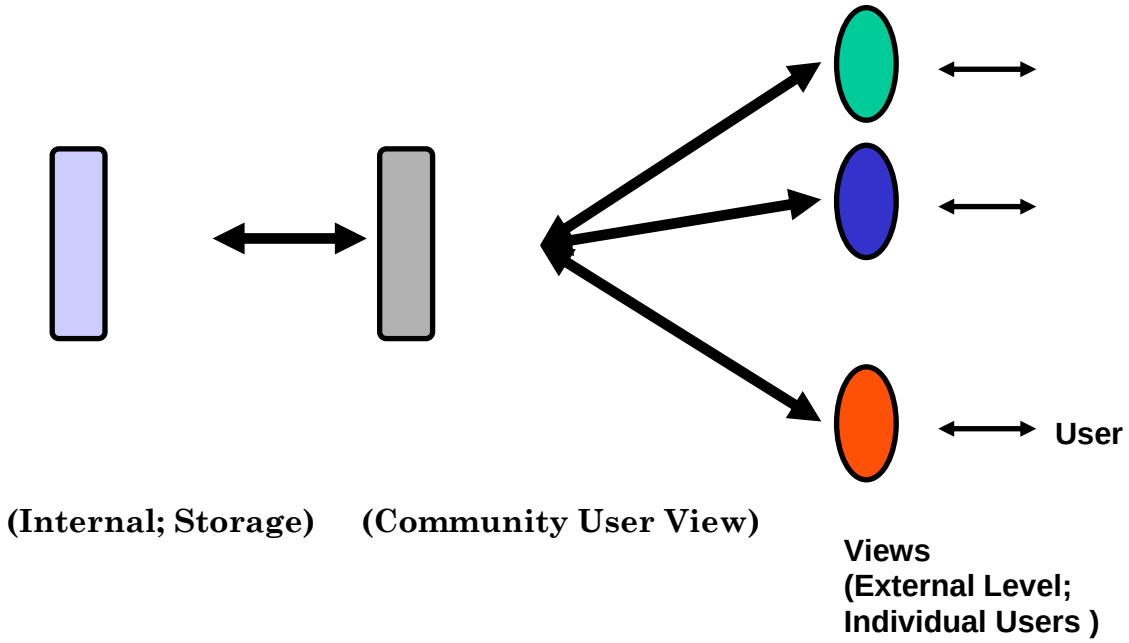

Figure 2: Three Database Levels

In a document, one defines the problem of finding information as a problem of finding some specific words or a collection of words. Since many words are not informative, the concept of *keywords* was introduced and extensively used. A keyword is a word that is considered rich in information content and is used by people for searching for information. Clearly words like a, an, the, is, are, etc are not very informative by themselves, but words like mercury, dog, god, and USA are considered relatively rich in information content. In general a keyword is an entity or object, place, or a concept. People will search documents related to, say, Abraham Lincoln and the system should point to all documents that are related to the keyword being searched.

In Figure 3, we show the basic architecture of a search approach. This general diagram could be applied to any search problem. Let us discuss information retrieval using some basic blocks from this architecture.

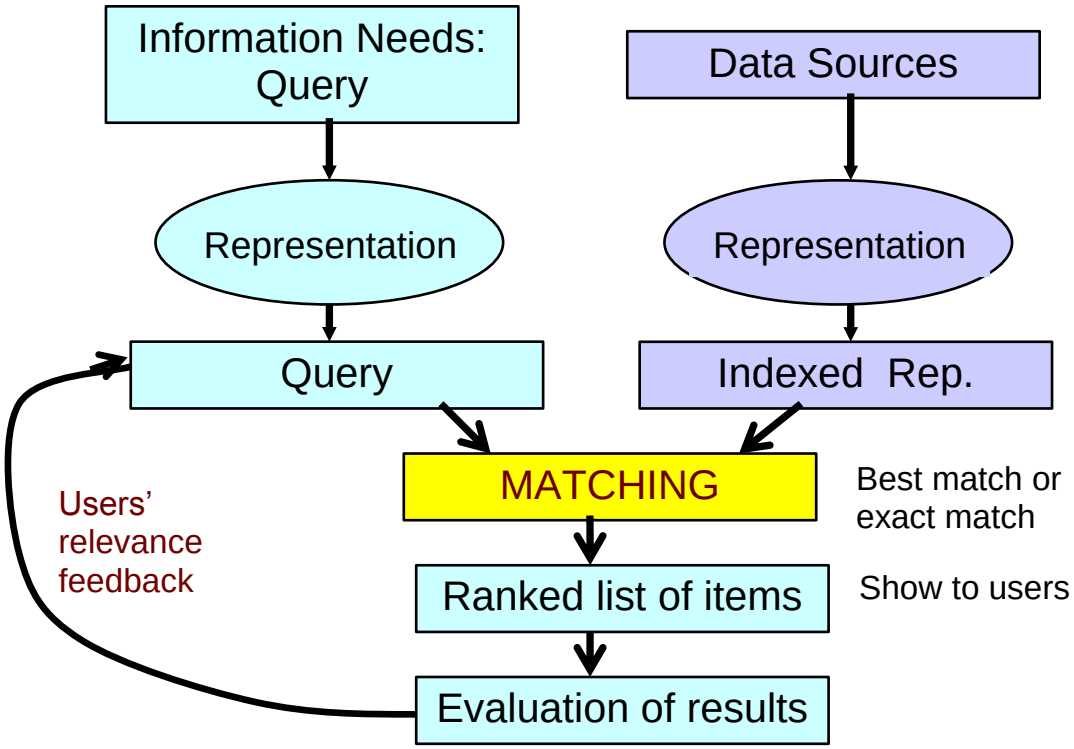

Figure 3: SearchArchitecture

In information retrieval data is in the form of a collection of text documents. In the collection there may be many documents. The information retrieval system should identify all documents that will be relevant in the context of a search. A data source in an IR system is a document that is in a file; on a web this could be a web page. Each data source should be represented in a form that will facilitate searching. As discussed above, structuring of data helps in organization and indexing resulting in efficient searches. Since text is not structured data (it is commonly considered semi-structured data), we should represent each file in a structured data form that could be indexed and searched. *Logical Representation* , commonly called a logical view, is used to represent the document or the original source using the data that captures the essence of the document from the perspective of search. One defines or designs a logical representation based on the types of searches that should be performed by users of the system. A logical representation is designed by considering:

- Attributes or features that can be used by users to define their problem, and
- The system can automatically extract those attributes and features from data sources.

One uses representation to index documents by essentially creating an indexed representation of these features and linking those to original documents. In IR systems, keywords are used as logical representation because

- Keywords are understood by people, and
- Keywords could be extracted using very simple text processing techniques.

A user articulates his information need in terms of a keyword or some combination of keywords. All queries must be articulated using the same logical representation as the system uses. In some cases, user's representation may not be exactly at the same level. In those cases, the system should translate user query to the IR system such that the IR system gets the query in same logical representation. The query in logical representation is matched with the indexed representation of the documents. The matching process could be very simple as just finding a simple term or could be very complex as we will see in multimedia cases in some of the following sections. Based on the result of matching, the IR system may find many documents that may satisfy the need of the user to a varying degree. Unlike database systems where search is binary, IR systems do not provide direct records that give users the answer, but they provide sources where the information could be found. Thus, the result of matching only indicates that a particular document may satisfy the user's information need.

The list of documents that satisfy user needs may be presented to the user. In many cases this list maybe long and it may not be a practical idea to present the complete list. In such cases, the list should be ranked based on the relevance of the document to the information need and only a subset of higher ranked documents should be presented to a user. A user may look at the list and may provide feedback to the system in terms of which documents are relevant to the information need and which are not. This information provided by the user, commonly called *relevance feedback*, may be used by the system to modify the query and reused to get new list from the system.

For evaluating the performance of a search system two commonly used measures, discussed in more details in a later section, are recall and precision. Recall characterizes how effective is the system in finding all relevant answers from among those who could be considered relevant in the whole document space. Precision is the measure of accuracy of results among all those that are presented as relevant answers. Precision gives us an idea of how effectively system distinguishes between correct and wrong answers.

Documents and Index

A document is not directly used in the matching operation. An IR system computes logical representation for each document and organizes an index of the logical representation of the documents. This index is used in matching. This is very important step because documents could be semi-structured or even unstructured but the system can structure the index and use it efficiently in search operation.
Inverted file index has become the most popular method to index documents based on their representation as keywords. An inverted file index is basically a list of words and all documents where this word appears. Thus, an index at the end of a book is a form of an inverted list.

Inverted files are very well suited for IR systems because the logical representation used in these systems for documents is keyword. Inverted files are very efficient in providing all documents that contain that word and its position in the document. For simple logical queries containing logical combination of two or more words, it is easy to compute logical combination of the documents that will satisfy the query.

Processing to prepare the Index

All documents in the system must be processed off-line to answer queries efficiently and in reasonable time. Since the logical representation of a document is keywords, the most important step in text processing is finding all keywords and their location in the document and storing this information in the inverted file index. The system finds all keywords by scanning the text and finding all words by detecting delimiters, including space, and then eliminating all articles, adjectives, and verbs. Dictionaries play a very important role in text processing because it is assumed that all valid words are in the dictionary.

Ranking Results

Since a query may result in many documents and the results are presented in a list, it is essential to decide in which order the documents should be listed in the results. Most people only see items at the top of the list. Ranking algorithms are developed to assign ranking to the result so that the list contains documents in decreasing value of their ranks. The rank of a document reflects its importance compared to other documents satisfying the query. The importance of a page is judged based on many factors such as

- Number of times the word appears in the document
- Position and fonts used for the word (in header, boldface, size)
- The popularity of the document as reflected by the link structure of the Web.

The final importance value is assigned using a weighted combination of factors

PageRank: Determining importance of a Document Node

A very important component in ranking results is the importance or popularity of the document (or an image or video as the case may be). To determine the importance or popularity of a page one may derive inspiration from a commonly used idea that is best manifested in academic circles in the forms of citations. One may consider citations and links in web pages similar in that they both point to a source that is considered relevant and important. A document, or a book, is considered more popular depending on Number of people referring to the document, and importance of people referring to the document.

If one could develop an approach that considers all links in all documents, in the case of the Web all pages, and develop an approach that could consider this *link graph* to assign a numerical value of importance to each page, then we could consider the numeric values representing the relative importance of each page. Let us understand this idea using a simple example shown in figure pagerank.

The Web has its link structure reflected by connections of each node to other nodes using links. Consider a node C. The importance of this node is determined by the incoming links to this node and the source nodes of those links. In this case the weights of the incoming nodes are 40 and 23 and are added to assign the node C the PageRank of 63. Since the node C points to 3 different nodes (D, E, and F), each link gets the weight of 21 by equally dividing the pagerank among all outgoing links from it.

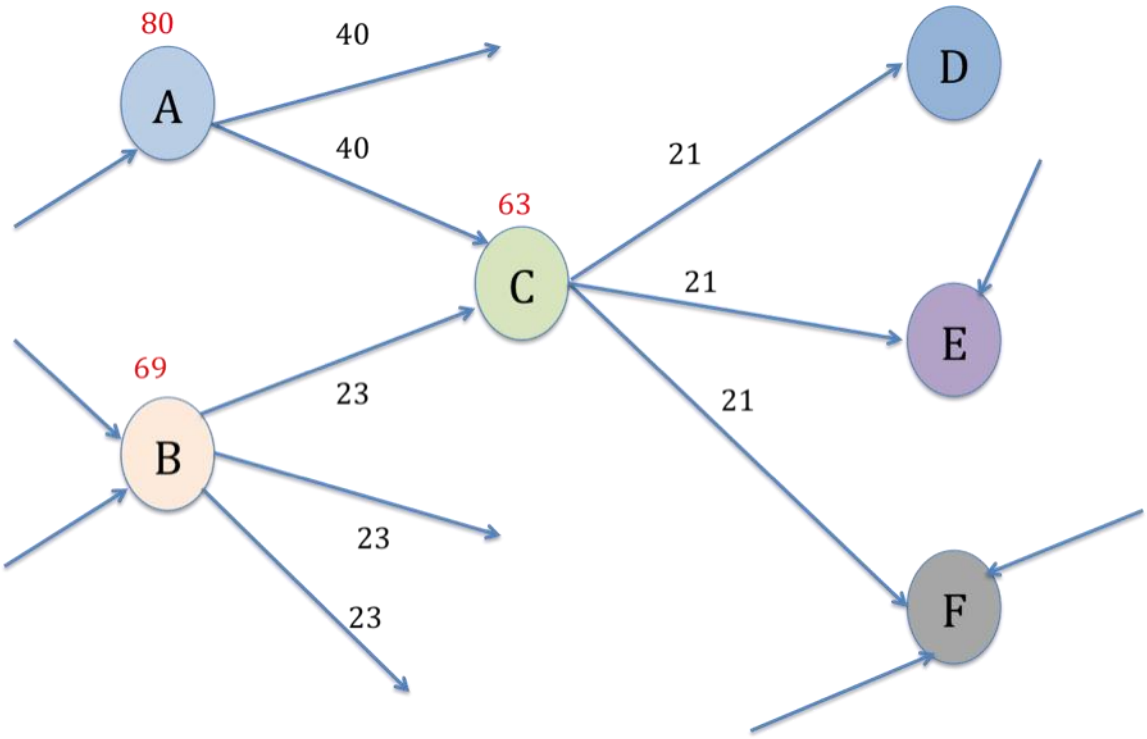

Figure 4: The pageRank of the node C is determined by the weights of the incoming edges from nodes A and B and is distributed equally among the nodes it points to.

This process of assigning pagerank to each node by summing up weights of all incoming links and distributing this among all outgoing is applied to all nodes on the Web, By creating an adjacency matrix to represent the link structure of the Web, and applying eigenvector analysis, computational approaches have been developed to compute the pagerank of each node on the Web. In formulating the problem, one needs to solve several pragmatic problems and mathematical issues that are discussed in detail in the literature.

Concept and Sentiment Search

Keywords are effective in many searches and have become very popular. They are limited in applications that require concept or sentiment search about a document. Suppose you were interested in finding all articles that show popularity of Michael Jackson. What keywords will you use? 'Popularity of Michael Jackson' will try to find

documents that will contain the terms popularity along with Michael Jackson. Clearly most documents that will say good things about him may not contain the term popularity.

Keywords are effective but are limited. Keywords could be considered the most basic representation to search for information in text documents. For more sophisticated searches, one must define abstractions on keywords that could be used by people. These abstractions could be some kind of 'models' that could be used to translate a user query into keywords that could then be used by the system.

One approach that has gained popularity because of its success in several applications is 'Bag of Words' model. In this representation, a concept is modeled using a set of words. In the above example, the term popularity may be replaced by a set of words such as attractive, beloved, famous, leading, likable, noted, pleasing, praised, prominent, social, sought, trendy, well-liked, and well-received. The system will then replace the keyword 'popular' by each of these and translate the query into multiple queries that will be used to find all documents and then the results will be assembled to present to the user. In a Bag of Words, one could also assign weights to different words to reflect their importance and use that weight while assembling the results.

Multimedia Information Retrieval

MIR has several important components. In Figure MIR, we show a general architecture of a MIR system. We can understand various issues related to MIR systems by considering the role and functionality of each major component as well as interactions among these components. A major challenge for building successful MIR systems is to understand not only each component but its role and interactions in the system. The main components of a MIR system are:

- **Media processing to extract features:** Sensors collect data, but all the processing is done based on information derived from this data, In some cases, the information is at the level of the application, but in most cases, one must rely on intermediate information, commonly called features. Features are usually a bridge between the data and the application. In every type of signal understanding, feature selection and feature detection is one of the most important steps. Computer vision, image and video understanding, speech understanding, and (even) text understanding are difficult problems that are actively exploring this area. Identification, efficient computation, and effective utilization of features are and will remain important research areas. Issues related to features and feature detection in different media data were discussed in Chapters X.

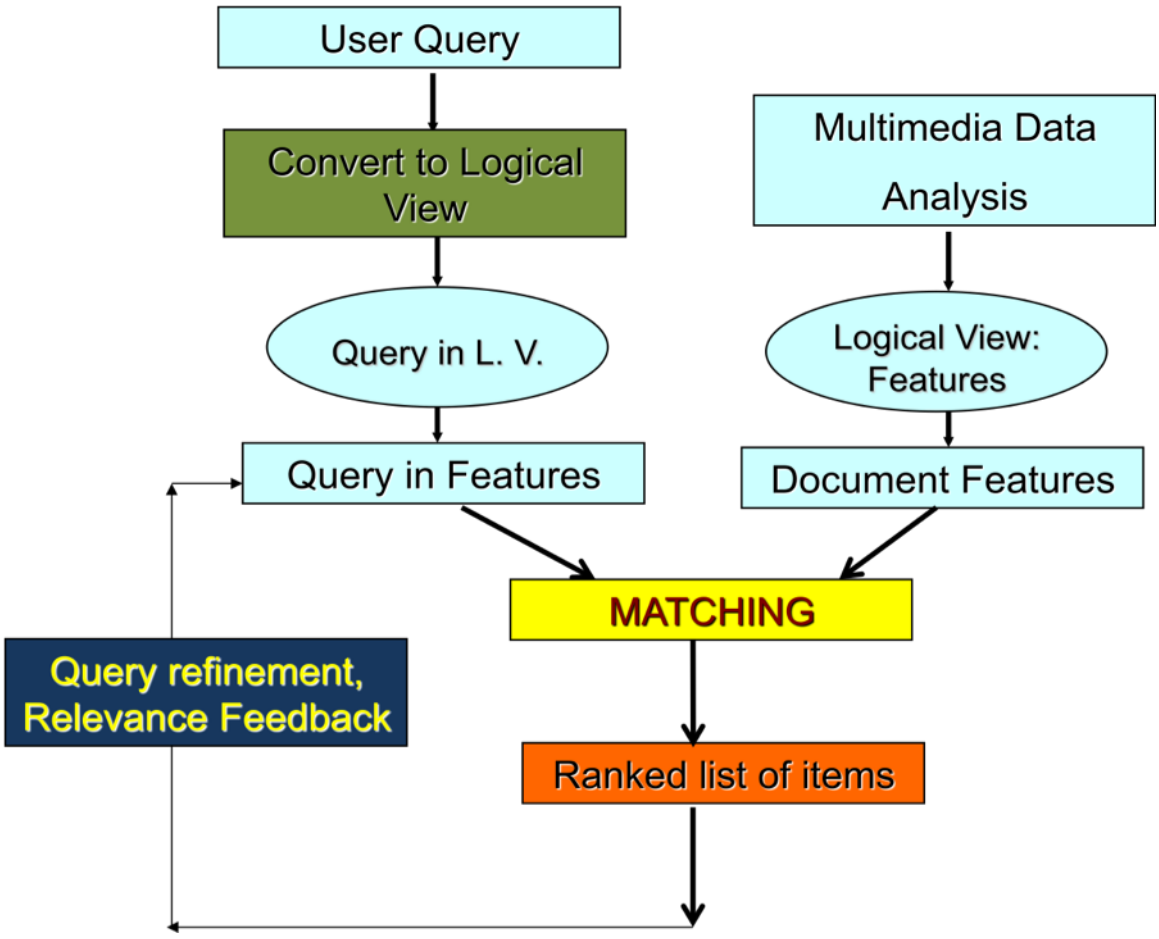

Figure 5: Multimedia Information Retrieval: High-level components.

- **Storage architecture of Media and Features:** Different types of media data have different structure and volumes. The media data should be stored along with its connection to features that are extracted from it and application knowledge and information that is derived from it. This multilevel linked structure should be stored to facilitate efficient and effective operations on the data and derived information. An additional factor is that the source of the data and locations where the features and application information is stored may be geographically separated. In fact, increasingly all these locations will be virtualized leading to architecture that will make everything unified for the users at different level.
- **Indexing:** Organization of information means finding suitable approaches to index so one can efficiently gain access to it. In multimedia data, this becomes a serious challenge. In most applications, one is not interested in the medium, but in the information that is derived usually as a combination of more than one mediums and sources. Current approaches to indexing are strongly influenced by the medium. Unified indexing approaches should allow accessing information of

multiple medium based on information requirement. Another problem is the dimensionality of data. Number of features required to characterize an image and other media data is very large and requires multidimensional indexing techniques.

- **Interaction environment:** The most common method to interact with text retrieval systems has been to formulate a query to the system and the system responds by providing the answer. The system is stateless – it does not remember your earlier queries and its answers are independent of the situation of the user. Increasingly systems are trying to introduce some knowledge about the state of the user and are also trying to personalize the responses to the user by using profiles and other information related to the user, as shown in Fig. MIR. Relevance feedback has been used as a mechanism to refine the response of the system for a specific query. Thus, though most systems still utilize a simple stateless query and answer system, many applications are starting to embed search or retrieval systems in their environment to make the whole environment more interactive and contextual. Many applications are naturally designed to use an incremental approach to solve a problem. In such systems, retrieval is a component that serves in the background by bringing in right information to the user at right instant.
- **Presentation and distribution of multimedia information:** Current search engines have adopted ranked lists as a standard approach to present the results. MIR is also using a variant of this. Considering that Multimedia data is not as natural to list based presentation as the text is, different other techniques are being explored. Increasing use of wireless mobile devices is making this a further challenging problem. Moreover, in case of live data in MIR, this problem may become further challenging.

Clearly, each component presents interesting challenges, but the real challenges start when one starts considering putting together real application of MIR. In the following, we will try to address important issues and direction in these areas.

Visual Information Retrieval

A visual information retrieval (VIR) system goes beyond text-based descriptors to elicit, store, and retrieve "imagery-based" information content in visual media. The basic premise behind VIR systems is that images and videos are first-class information-bearing entities and that users should be able to query their content as easily as they query textual documents, without necessarily using manual annotation. The domain of VIR has inherited the analysis component of computer vision and the query component of databases and information retrieval systems by tapping older disciplines of computer science: database management and information retrieval systems and image processing and computer vision. To understand VIR issues and techniques, we should address three basic questions:

- What constitutes the "information content" of an image or video in the specific context of any application? Or, what is visual information?
- How can a user specify a search for a desired piece of information?
- How efficient and accurate is the retrieval process?

What Is Visual Information?

Two kinds of information are associated with a visual object (image or video): information about the object, called its metadata or context, and information contained within the object, called content and represented by visual features. Metadata is alphanumeric and generally expressible as a schema of a relational or object-oriented database or using XML. Visual features are derived through computational processes—typically image processing, computer vision, and computational geometric routines— executed on the visual object. The simplest visual features that can be computed are based on pixel values of raw data, and several early image database systems used pixels as the basis of their data models. These systems can answer such queries as:

• Find all images in which at least 500 pixels are in the color range represented by (red = 240 to 255, green = 130 to 200, and blue= 0 to 30).

• Find all images that have about the same color in the top half region of the image as this particular one.
• Find all images that are rotated versions of this particular image.

As can be seen, above queries are based on image content but are not related to any objects or concepts. If the user's requirements are satisfied with these, data modeling for visual information is almost trivially simple. However, a pixel-based model suffers from several drawbacks. First, it is very sensitive to noise, and therefore a couple of noise pixels may be sufficient to cause it to discard a candidate image for the first two queries. Second, translation and rotation invariance are often desirable properties for images. Third, apart from noise, variations in illumination and other imaging conditions affect pixel values drastically, leading to incorrect query results.

These limitations are not to say that pixel-oriented models are not used in visual information retrieval. Significant image and video segmentation requires considering and analyzing pixel attributes. However, multimedia information retrieval based only on pixel values is not very effective because as shown in Fig Semantic Gap, humans usually ask queries in terms of objects and events, rather than pixel attributes..

Image Retrieval

Content Based Image Retrieval or Image Similarity

Early image retrieval techniques were developed to consider image collections. These techniques were called 'content based' retrieval because they considered image characteristics for retrieving images. An image could be represented using its basic features. Commonly used basic features are
Color

Texture
Structure or shape

We discuss these features in more details later in chapter X. These three attributes of an image are captured using different types of image characteristics computed using image processing techniques applied to the pixel characteristics.

Early image retrieval systems were based on determining image similarity using color, texture, and structure features. Since these systems used characteristics of image content, they were commonly called Content Based Image Retrieval (CBIR) systems. The very early versions used a weighted combination of the above features to judge image similarity and rank pictures based on these. Sometimes these systems were also called Query by Image Example because the query for an image was another image.

Figure 6 and Figure 7 show two examples of these systems. In both these figures, screenshots for a query and its results are shown. In both figures, the query image is the first image, the left-top image and all images are shown in the order of their similarity to the query image from the database. The number of images in the database for these examples was 20,000.

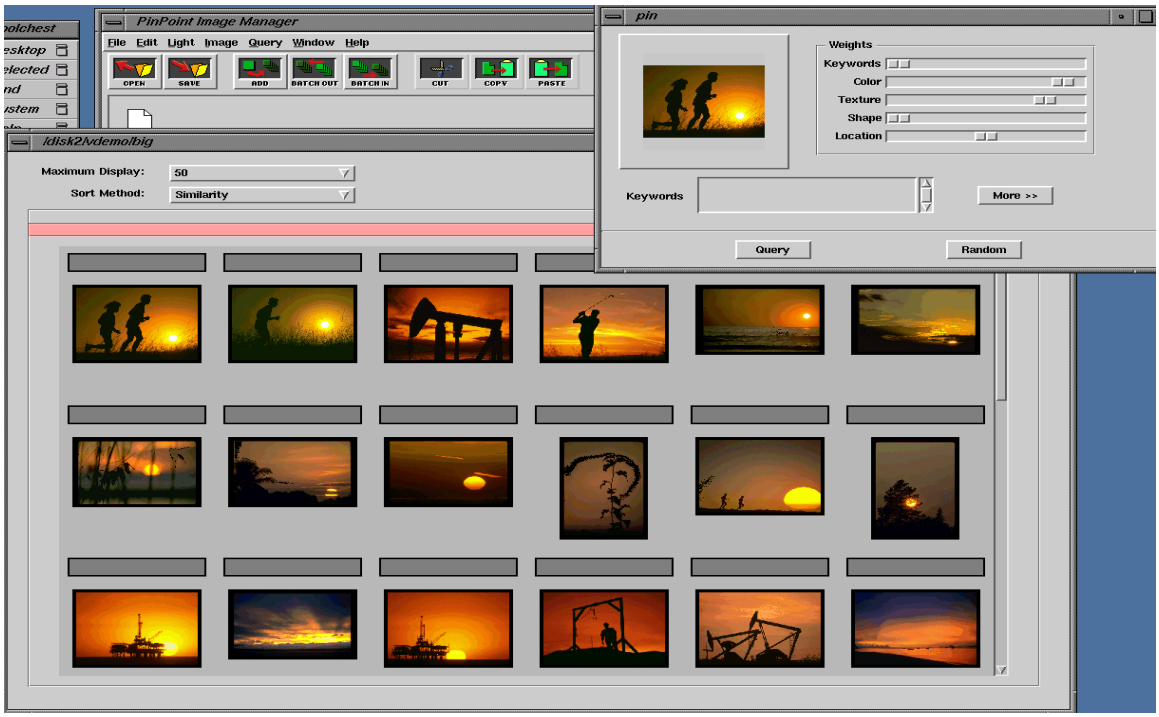

Figure 6: A screenshot showing the selected image, the weights for different features and results corresponding to this query.

Figure 7: For the query related to a rose, in the top left corner as the first result, the top 50 results are shown here.

In Figure 6, at the right top the panel shows that a user could assign importance to different image features in the query.

A very important point to remember about these systems is that these systems are ranking images based on similarity of visual characteristics of images. These systems do not have any concept of objects as we know them. Thus in Figure 7 one can see that most images are similar to the query image because they contain large number of green pixels with similar textures and in the centre is another object of different color – in most cases similar to the color of the query image. The system is not trying to detect roses as in the query image, though majority of results contain flowers, particularly roses.

Many enhancements have been made to these basic techniques. One can divide each image into multiple regions, say a 3 X 3 grid and computer these features for each region and compare these with the corresponding ones. The similarity is then judge based on how close each region is to corresponding region. Another simple modification is to consider color histogram prepared using only coherent colors. This is done by ignoring those pixels whose color value is different from its neighbors. This means that most noisy pixels and pixels on edges are ignored and only those pixels in homogeneous color regions are considered.

Searching Based on the Semantic Content

If content of images in terms of objects in it and relationships among them is somehow available, then it is possible to develop powerful retrieval techniques that could be called content based retrieval. Approaches discussed in the previous section were based on computing average attributes of pixels in a query image and using those to retrieve similar images based on these attributes, In more realistic situations, one may consider following queries:

- Show all photos of Jay in which he is participating in a sporting activity.
- Show me all photos of Jay with Tarah in Cabo.
- Was Tarah with Jay in Paris?

These queries require recognition of objects and their relationships in photos. These queries cannot be answered without clear object models.

One may also consider following scenarios in which a camera, say a mobile phone camera, is on and the following query is issued:

- What kind of insect is this?
- Is this plant healthy?
- Who is this person?

Above queries also require recognizing objects in the field of view of the camera but are more like query by example as shown in above section.

Recognition and Annotation

Object recognition has been a very important research area in computer vision for about 50 years. In fact, a more general research area is pattern recognition that deals with identifying patterns that may correspond to objects, concepts, or activities in data. The fundamental problem addressed in pattern recognition is: Given N patterns (P1, P2, …,Pn) that are models of corresponding objects or concepts; identify which of these patterns are present in a given data set.

Clearly recognition and retrieval are not the same problem, but recognition may play a very important role in content-based retrieval. The queries listed above all require recognition of objects and concepts and relationships among them. This important fact has resulted in application of several recognition techniques to detect objects and annotate images with these objects for retrieval purpose.

Since recognition requires models of objects, defined above as patterns, the first step in developing recognition approaches is to have strong models for objects. Considering the variability of objects, there appearances, and changes in appearances in images due to different viewpoints, illumination conditions, and climatic conditions it is very difficult to create models for objects that could be used for recognition in images. Popularity of

machine learning techniques in computer vision and multimedia is primarily for simplifying the process of creation of these models that could be used in recognition.

Despite significant efforts in development of automatic recognition techniques, the progress in this area has been quite slow. Since the need for organization and retrieval of images has become a real hurdle in the growth of many applications, manual and semiautomatic approaches received significantly popular.

A very popular concept to emerge in many applications has been that of 'tags' for images and video. This concept has been used in many other applications, including in text documents. A person could look at an image and assign it tags that will describe it. The tags could be list of objects, concepts, or names of objects or places. In fact it could be anything that could describe the image and help in management and retrieval of the image. Due to increasing use of images in many applications, particularly in image sharing, use of these tags appeared quite attractive.

Some common problems with manual assignment of tags that must be considered in any tag based retrieval system are:

- Tags are very subjective. A picture is usually assigned different tags by different people.
- Tags are time dependent. Depending on when tags are assigned to a picture the tags could be very different.
- Availability of tags for pictures is random. For most photos people do not assign tags. It is commonly observed that less than 2% photos on most photo sharing sites have tags. Moreover, people do not assign tags to most photos in their own collections.

Image Search on the Web

Most images on the Web in early stages appeared as part of documents. In these cases the images could be considered secondary to the text material because images were used more to provide experiential component in an otherwise text document. Searching for images on the Web started out very similar to text search. Images are searched on the Web using keywords. The keywords that characterize images are usually from the following:

- Name of the image file,
- Text on the page where the image is located, or
- Tags assigned to the image.

In early systems for Web Image Search, image files were not even opened to analyze them. An image file was detected from the extension in the file name and then the search was performed using text. This approach does have limitations, but fits well with techniques used in text search.

In the last decade, many applications emerged where photos or images are the main part of the document. In these applications, there could be text that is used as the supporting material, but the photos are considered the main carrier of information and experiences. In such applications, the techniques based on text analysis and keywords are inadequate. In some of such applications, tags have been used, but as mentioned above, the limitations of tags make them only partially useful.

Video Retrieval

Video is more than just a sequence of images; it has synchronized audio also. Depending on an application, the information content in a video may be more in audio, particularly speech, more in images or equal in both audio and images. In any case, video must be considered very differently from images.

Information content in a video depends on the type of video. Most videos could be considered in one of the following classes:

- Produced Video: Videos such as TV and Movies are produced by professionals following well defined conventions. These videos can be compared to well authored text where there is a well defined structure in document starting from a book to chapters to sub-chapters to paragraphs and to sentences. Produced videos also have such well defined structure.
- Semi-produced videos: Videos such as sports and seminar lectures are semi-produced. They also follow some general conventions, but these are not so rigid.

Amateur Video (Unstructured videos): These are the commonly produced video by most people. Using a video camera people may collect many video segments and may edit them using one of the commonly available video editing systems. Most of these videos do not follow any videography rules.

Video Segmentation

Video segmentation is used in two very different senses. In computer vision and some related fields, video segmentation is used to mean detecting meaningful objects in video. Thus, a car or a person in a sequence must be segmented based on its appearance and motion despite the fact that this object may look very different in different frames of the video. Many techniques, including motion detection, tracking, and structure from motion among others are used for this purpose. This has been a very active research area now for several decades.

In multimedia information retrieval techniques, the above techniques play important role, but the term video segmentation is used to represent structure of video, rather than content and activities in the scene being captured by the video. This structure is defined following concepts developed in video production community. This structure is shown in Fig VideoSemenatation.

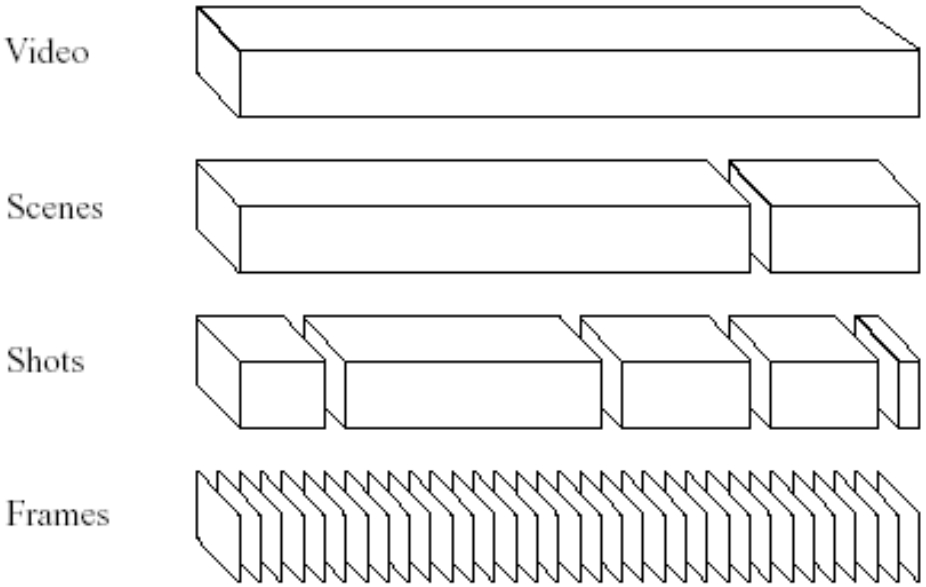

Figure 1: Standard video structuring model

Figure 8: Video could be segmented starting with each frame as the basic unit and then grouping into shots, scenes, and episodes.

We consider video segmentation in the sense of parsing a video rather than partitioning a scene in its individual components. The atomic unit in a video is considered an image frame. A video contains these frames acquired at a regular interval. In acquiring video, the camera is either at a stationary position or in uniform motion to acquire a shot. The shot is considered as a unit of interest in video for building stories. There are many types of shots [Ref] that are used by videographers. In video segmentation, the goal is to group all frames that may form a shot. Many approaches have been developed for automatic shot detection in video. Such techniques were discussed in Chapter X.

A shot in a video is usually represented using one or more 'key frames'. A key frame is simply a representative frame for the shot. For shots of long duration more than one key frame could be used. The next step in the segmentation is called scene detection. A scene in a video is considered a group of shots that could be considered related to an event or an episode. Thus if two people are discussing then even though the shots may represent alternating between two people, until the conversation is going on, the shots (and the frames belonging to those shots) could be considered to form a scene.

As can be noted, the above segmentation of a video may be considered similar to the notions of words, sentences, paragraphs and chapters in text. The definition is not very rigid except that a frame is definitely a very well defined unit as is a word in text. After that the analogy does not work very well.

Video Retrieval

Suppose that we have a collection of video. The types of queries that people may ask may be:

- Show me all 3-pointers by Kobe Bryant.
- Show me 3-pointers by Kobe Bryant in the last 5 minutes of the game.
- How did BP finally stop the gushing oil?
- In how many parties, was Tarah wearing Pink dress?
- Show me all romantic scenes from Titanic.
- Show me all stunts by Tom Cruise.
- Show me the car stunts from Casino Royale? What car was Bonds driving?
- List all popular romantic movies in the last decade.
- Show me all players who 'pull' like Tendulkar.
- Show me all stories on BBC that talked about the Psychic Octopus.

All these queries require that the video is analyzed and some information is extracted from video. The queries may be based on

- Objects
- Activities
- Time of activity
- Similarity of activity to a given activity
- Genre of activity
- Visual attributes of objects
- Concepts

And this may require analysis of visual characteristics, audio, and meta-data related to the video.

Many Faces of Video

Video has rapidly emerged as a dominant medium for use in diverse applications. Its use continues to increase rapidly with popularity of smart phones. A major advantage of video is that it can be easily adopted to work for people from different language groups either through sub-titles or through dubbing. Another major factor in popularity of video is increasing use of computing in Television. With much of the TV being supplied using

cable and satellite services, the home receiver box, commonly called the set-top box, has become a computer that can store, process, and serve video more like a computer than the traditional TV. This is resulting in convergence of computing and Television, commonly called IPTV.

Traditionally TV was viewed as a passive medium in which a user only selected a channel, and changed it occasionally, and just consumed the video that was being broadcast. Increasingly TV is becoming more like a database of video that can be accessed by a user as she wishes. A user may be passive as in early days of TV or could be very active as video-game generation. Another user may even want to cut and paste parts of videos from different sources and compose his own playlist. Such a playlist may even be automatically generated.

Multimedia Information Retrieval

Indexing Single Media or Multimedia using events

Multimedia data has disparate nature with respect to what it represents in terms of physical attributes over space and time. It is captured using different devices and to make sense, it must be rendered using different devices. The fundamental difference in the nature and volume of data results in the storage of data using different file systems. The capture and rendering technology for each type of media also results in different type of organization of data to match human perceptual perspective of that mode. This results in different approaches to indexing of the data. For example, the video maybe indexed based on the time code, while photos maybe indexed based on photo number and text as page and line numbers. These indexes are fundamentally different and are incompatible. This creates silos among different types of multimedia data. This situation is shown in Figure 9.

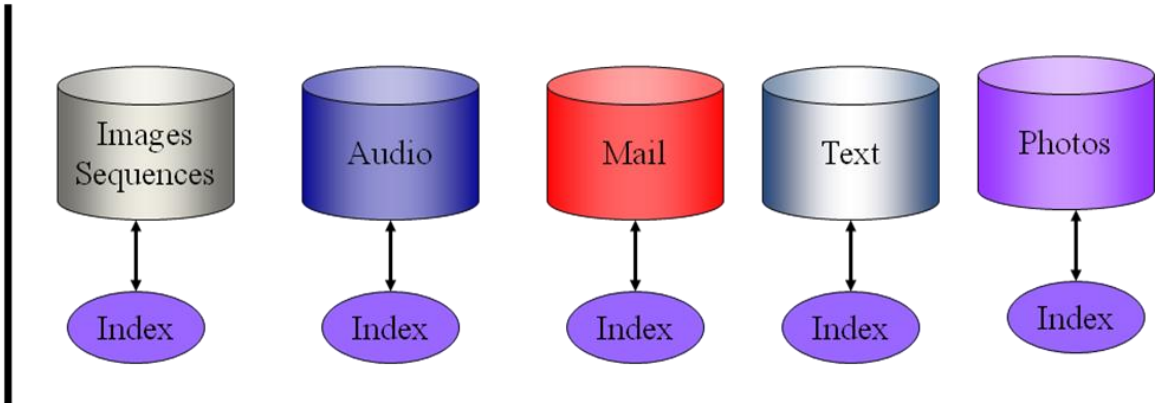

Figure 9: Different type of data and their indexes create isolated silos of multimedia data.

A challenging question for the MIR research is how to break these silos to unify multimedia data? Multimedia data represents complementary information about the world captured using different types of sensors. Since the fundamental reason for multiplicity of data is multiplicity of attributes it represents, it is natural that this data has different characteristics in terms of its storage and indexing requirements.

The only unifying basis behind this data is not in the data but is in the real world. This data is collected for something in the real world that is considered important. Objects and events are complementary in modeling the real world around us. Multimedia, due to its temporal nature, is particularly important in dealing with events. In fact, different modalities of data capture different aspects of events and objects. Since different streams of multimedia data have nothing in common except that they are all related to the same real world event, it is natural to index them using events and their parameters. This is an effective way to unify and index the multimedia data resulting in the situation shown in figure 10. By defining an event model to represent all essential aspects of an event as well as all its associated multimedia, it is possible to create a unified approach to indexing multimedia data.

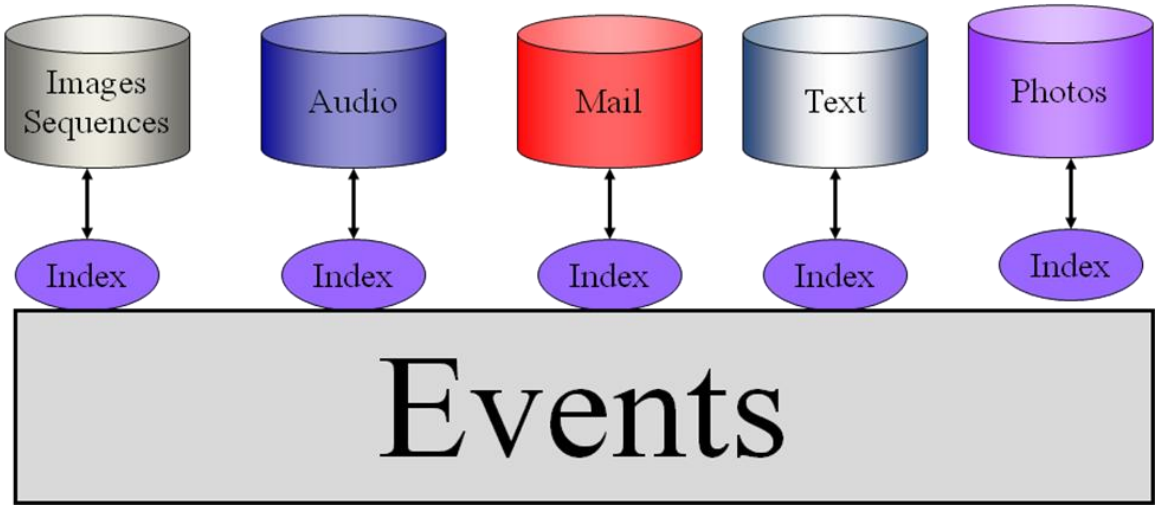

Figure 10: Real world events could be used to unify different media and provide a unified framework for indexing multimedia data.

Relevance Feedback

In many cases, articulation of a query precisely to get answers from an information retrieval system is not easy. This is because use may have only a vague idea about what she is looking for, or the concept or information sought may be difficult to precisely articulate. For example, a user may be looking for people who look like Abraham Lincoln. How can one articulate a query to an image retrieval system to find all people who look like Abraham Lincoln? To deal with such situations, the concept of successive refinement of queries later popularized as relevance feedback was introduced. As shown in Figure 11, based on the query by a user, the system produces ranked list of results.

The user then can provide explicit or implicit feedback to the system by conveying which of the results listed are not relevant to the query posed by the user. The system can then modify the query and fetch a new set of results.

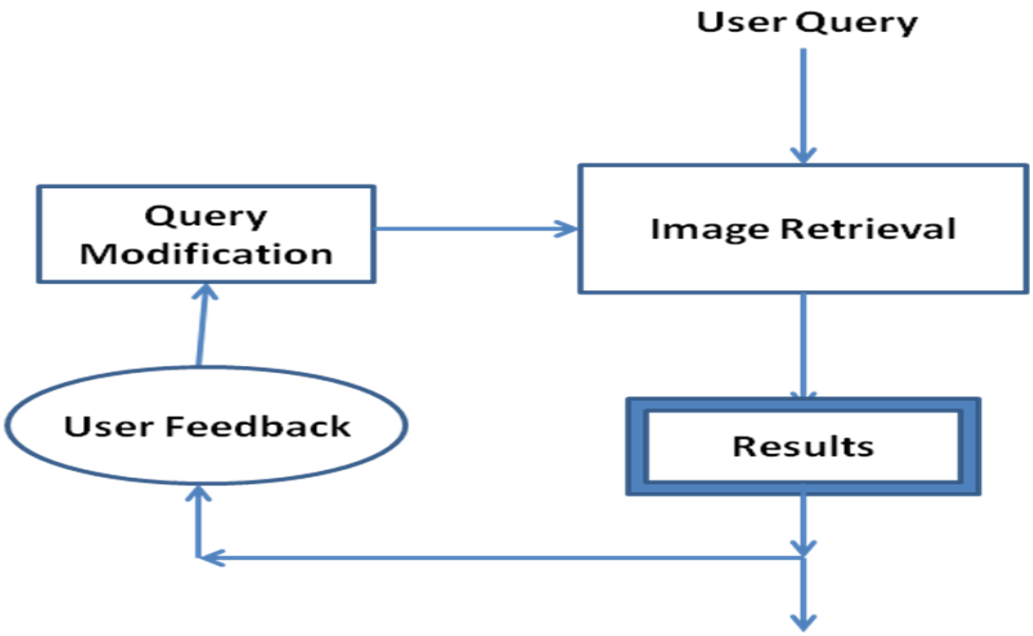

Figure 11: A user query may not represent what the user really has in mind. An iterative query refinement may be used to refine the results to suit user needs.

A simple way of looking at this approach is to consider this as a simple classification problem, as shown in Figure 12. Based on the query by a user, the system considers all documents, or images, in two classes: relevant documents and not-relevant documents. The classification is done using a mathematical approach that considers different features. If the user is not satisfied with the results, then the classification function used by the user and by the system are considered different. The goal of the relevance feedback is to get more information from the user about the classification function that the user has in mind. Based on the feedback from the user, the system changes its classification function to align it better with that of the user.

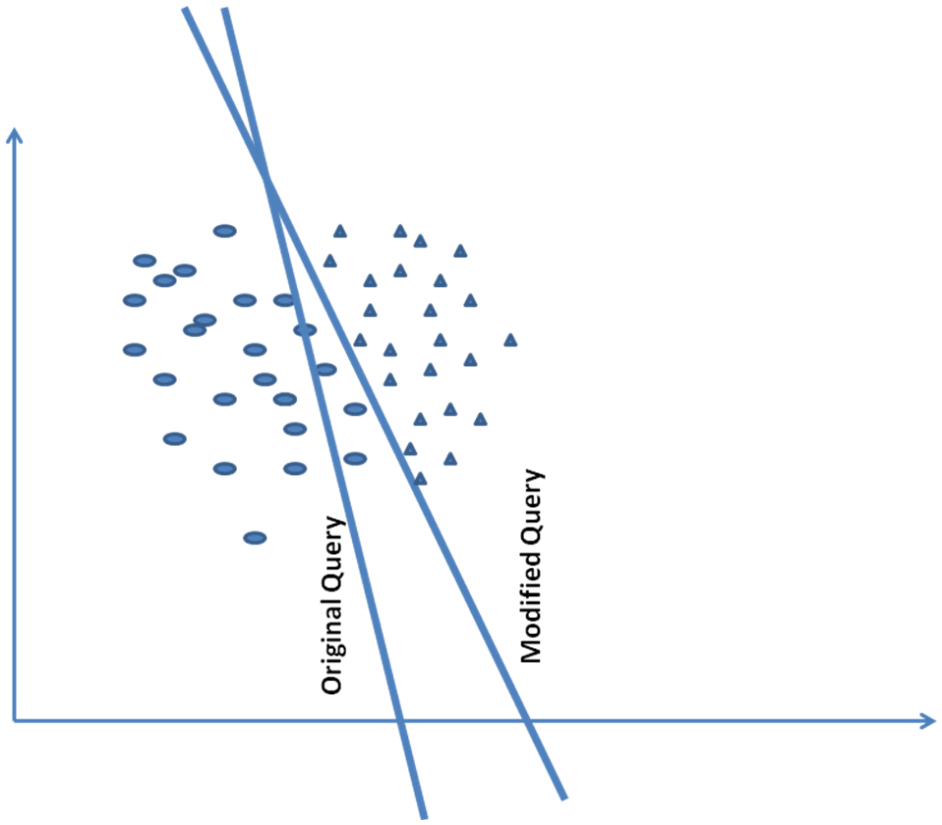

Figure 12: Based on the user query, the system uses a classification function that gives unsatisfactory results. User's feedback about relevance results in a modified query that gives better results.

The above discussion using classification function works very similarly if in place of classification function, we consider a ranking function.

Summaries and Storytelling

Number of photos, videos, audios, and text descriptions of events has increased very rapidly such that the sheer volume of even personal collection of photos has become tedious to manage and maintain. In this section we discuss two related but different problems. Though we will be discussing using more examples from photos, but the discussion will be equally applicable to other media and to the combination of these.

Summarization

As in text documents, with multimedia content also, a summary represents a condensed version of the information contained therein. Consider personal photos. Suppose that you go on a trip to Brazil and comeback with 3000 photos that you took. You want to share those photos with your friends and family. Would you show them all 3000 photos? Or consider a similar situation. In year 2015 due to easy availability of digital camera, storage, and ease in transferring those photos to your collection, you take 25000 photos.

At the end of the year, you want to send a 'Year in Photos' to your mother, your significant other, your best friend from childhood, your professional friends, and your cousins. How do you select 25 photos to represent your year in photos? These are just two simple cases of summarization of photos.

In simple terms, photo summarization can de described as: Given a set containing N photos, how can one select M (where M is much less than N) photos that represent the set the best. One may consider summarization of photos as a semantic sampling problem such that the goal is to sample the photos, by selecting particular photos, that allow representation of semantics of the complete set satisfactorily. Unlike the sampling problem in signal processing there are two differences here. First, the photos themselves represent sampling of the event by the photographer, and it is difficult to capture semantics quantitatively to evaluate whether sampling is adequate or not.

Given the importance of the problem, many approaches are being proposed for summarization of photos.

Storytelling

One of the most popular and frequent art in human society has been storytelling. From oral traditions to sophisticated video production and now multimedia environments, people have used all possible technology to enhance communication of their experiences. With advances in multimedia, storytelling has taken new directions. We discussed several tools that one may use in story telling in the chapter on multimedia authoring tools. In this section we discuss techniques that will assist people in storytelling using the media that they collect.

A storytelling approach could be described as the one shown in Figure 13.

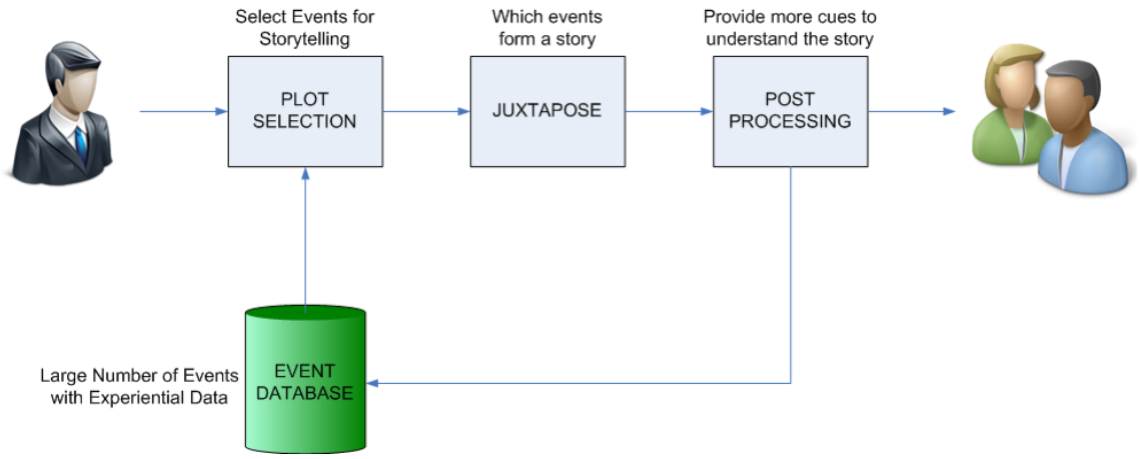

Figure 13: In a storytelling system, an event or a sequence of events is presented to audience using all data that is available. The main steps in such a system, manual or automatic, are shown in this figure.

Storytelling is based on a database of events and the experiences associated with events. A storyteller usually has a message or a perspective that she wants to convey using appropriate events and experiences associated with those. Another important factor in storytelling is the listener or the audience of the story. Events and experiences are selected based on the audience interest and profile also. Essentially, one may define,the Storytelling: given a set of events, and associated media, select appropriate events and for each event select appropriate media data that will maximize the message communication from the storyteller to the audience. This becomes a two stage process. In the first stage one selects appropriate events based on the message and then constructs a coherent story considering the profile and the level of audience.

Further Readings

Multimedia information retrieval is currently (in 2010) one of the most active research areas in multimedia. The growth of content in multimedia requires that tools to organize and index the content for rapid and relevant access must be developed. Starting in early 1990s, this field has grown substantially. Earliest concepts in this area were introduced in [5]. Research in query by pictorial example [21] started a trend that later became known as query by content. [Despite a vibrant research community and very active pursuit, the problems have been challenging. The most fundamental issue in organizing and accessing content is the semantic gap. Semantic gap was first discussed in [9] and later popularized in [10]. Many research papers are now addressing semantic gap. Many review papers have been written in this field. Interested readers are advised to start with one of these to get a good feel of the field. A particularly influential paper seems to be [10] that summarized research until year 2000 and is one of the most cited papers. Some later review papers [18, 19, 20] give a good summary of MIR research and are good complements to the earlier review [10].

Query by successive refinement was introduced in [3] and later was formalized in general cases using the concepts of relevance feedback in [17]. The concept of emergent semantics was introduced in [4] and is likely to attract more attention as the data size is increasing exponentially. Some trends to organize personal photos using context are explored in [2,6,7,13]. A good summary of metadata associated with photos is in [11] and how to use those is explored in [1]. Some researchers have started exploring concepts of visual words [12]. One must be very careful in using words in multimedia because words in text are manually delimited and are used as defined in a dictionary. Both these are not currently valid for visual words.

Use of events for unified access to multimedia was presented in [23] using models of events defined in [22].

References

1. Sinha, P., Jain, R.: Semantics In Digital Photos: A Contenxtual Analysis. In: Proceedings of the 2008 IEEE International Conference on Semantic Computing, IEEE Computer Society (2008) 58–65
2. Anguera, X.; Xu, J.; and Oliver, N.: Multimodal photo annotation and retrieval on mobile phones, Proceeding of the 1st ACM international conference on Multimedia information retrieval, 2009.
3. J. Bach, S. Paul, and R. Jain, "An Interactive Image Management System for Face Information Retrieval," *IEEE Transactions on Knowledge and Data Engineering, Special Section on Multimedia Information* Systems. Publication. 1993.
4. Simone Santini, Amarnath Gupta, and Ramesh Jain "Emergent semantics through interaction in Image Databases" IEEE Transactions on Knowledge and Data Engineering, summer 2001.
5. A. Gupta, T. Weymouth, and R. Jain, "Semantic Queries with Pictures, The VIMSYS Model," *Proceedings of VLDB'91, 17th International Conference* on *Very Large Data Bases*, Barcelona, Spain. Sept. 3-6, 1991
6. Bo Gong and Ramesh Jain, "Segmenting Photo Streams in Events Based on Optical Metadata", Proc. IEEE Conf. on Semantic Computing, Irvine, CA, Sept. 17-19, 2007.
7. Ling Liu and Tamer M. Özsu (Eds.) (2009). "Encyclopedia of Database Systems, 4100 p. 60 illus. ISBN 978-0-387-49616-0.
8. Sinha, P., Jain, R.: Classification and annotation of digital photos using optical context data. In: Proceedings of the 2008 international conference on Content-based image and video retrieval, ACM (2008) 309–318
9. Santini, S. and Jain, R. "Beyond Query by Example; IEEE Second Workshop on Multimedia Signal Processing, Redondo Beach, CA , **pp.** 3 – 8, USA 7-9 Dec 1998.
10. **Content-Based Image Retrieval at the End of the Early Years**
Found in: IEEE Transactions on Pattern Analysis and Machine Intelligence
Arnold Smeulders , et. al., December 2000.
11. EXIF: Exchangeable image file format for digital cameras: Exif version 2.2. Technical report, Japan Electronics and Information Technology Industries Association, 2002.
12. K. Barnard and D. Forsyth. Learning the semantics of words and pictures. In *Proc of IEEE Intl Conf Computer Vision*, volume 2, pages 408–415, 2000.
13. M. Boutell and J. Luo. Bayesian fusion of camera metadata cues in semantic scene classification. In *Proc. IEEE CVPR*, 2004.
14. M. Boutell and J. Luo. Photo classification by integrating image content and camera metadata. In *Proceedings of ICPR*, 2004.
15. M. L. KHERFI AND D. ZIOU AND M. BERNARDI **Image Retrieval From the World Wide Web: Issues, Techniques, and Systems**
16. GUPTA, A. AND JAIN, R. 1997. Visual information retrieval. Commun. ACM 40, 5, 70–79.

17. RUI, Y., HUANG, T. S., ORTEGA, M., AND MEHROTRA, S. 1998. Relevance feedback: A power tool in interactive content-based image retrieval. IEEE Trans. Circ. Syst. Video Technol. 8, 5, 644–655.
18. M. S. Kankanhalli and Y. Rui, "Application Potential of Multimedia Information Retrieval", Proc. IEE, April 2008.
19. R Datta, D Joshi, J Li, and J. Wang, "Image Retrieval: Ideas, Influences, and Trends of the New Age", ACM Computing Surveys, VOl 40, No. 2, April 2008.
20. M. L Kherfi, D. Ziou, and A. Bernardi, "Image Retrieval from the World Wide Web", ACM Computing Surveys, VOl 36, No. 1, March 2004
21. M Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yonker, "Query by Image and Video Content: The QBIC System", IEEE Computer, Sept. 1995.
22. G. Utz Westermann and Ramesh Jain," Towards a Common Event Model for Multimedia Applications", in IEEE Multimedia, January 2007.
23. Ramesh Jain, "Out of the Box Data Engineering: Events in Heterogeneous Data", Key note talk in Proceedings of International Conference on Data Engineering, Bangalore, India, March 2003.
1.

